美团网
meituan.com

数据平台治理实践

徐章

大纲

- 数据平台现状
- 数据平台治理服务
  - 资源管理
  - 数据质量
  - 数据安全
- 数据治理系统机制
- 未来的挑战
- QA

数据平台现状

- 数据仓库演进三个阶段
- 演化的动力与策略
- 服务角色转变
- 面临的问题

数据仓库模式的三个阶段

美团网 meituan.com

	封闭	共享	自治
时间段	~ 2013.3	2013.4 ~ 2014.12	2015.1 ~ now
模式	数据组支撑 所有业务	业务组参与开发 数据组监督规范	业务组独立建设 数据治理系统化
数据开发者	15~人	164人(季度)	413人(季度)
任务数	810	4700	14500
业务线	10~	23	32

演化的动力与策略

美团网
meituan.com

动力:效率

策略:开放,规范执行系统化

服务角色转变

美团网 meituan.com

前	后
面向业务需求	面向数据开发者
关注数据本身	维护平台
产出数据为目标	提升效能为目标

分治模式面临的问题

美团网 meituan.com

- 如何保证各业务组间的资源使用互不影响
- 如何监督不合理的资源使用
- 如何提升数据质量
- 如何保障数据安全

大纲

- 数据平台现状
- 数据平台治理服务
  - 资源管理
  - 数据质量
  - 数据安全
- 数据治理系统机制
- 未来的挑战
- QA

资源管理

美团网
meituan.com

- 隔离性
- 分配体系
- 资源使用有效性监控

隔离性 - 要求

- 业务线间互不影响
- 降低各数据使用场景之间的干扰

隔离性 - 策略

- 按业务组划分计算与存储资源
- 按不同的数据使用场景划分资源队列
- 定制标记每个任务所属业务组和使用场景的策略:
  - 生产: 根据任务写入目标库确定所属业务组,再根据业务组确定生产资源队列
  - 查询: 由查询发起者所在业务组确定查询资源队列

分配

- 资源单位定义:
  - VCore
  - Memory
- 分配与预估策略: 用户提预算+平台整体统筹

场景划分

- 对主要的使用场景定制分配策略
- 分配方案区分夜间和白天

主要场景	生产	开发测试	查询
资源量比重	夜间为主 白天为辅	较少	白天为主
优先级	高	中	低
分配模式	固定 可预期	固定	可伸缩

资源使用有效性监控

- 机制: 发现->通知->反馈->操作
- 策略:
  - DLM(数据生命周期管理): 数据热度分析:
    - 发现无用的计算和存储
  - DCRM(计算资源管理): 收集统计任务/分组级别的资源使用信息:
    - 发现性能异常的任务
    - 发现影响整体生产进度的风险
    - 资源使用趋势作为资源预估的依据

数据质量

美团网 meituan.com

- 数据质量关注点
- 质量保证策略
- 数据质量评估

数据质量特征

- 时效性
- 准确性
- 完备性
- 一致性
- 可用性

生产过程中的质量关注点

美团网 meituan.com

	时效性	准确性	完备性	一致性	可用性
ODS		应用埋点逻辑正确,上游数据生成无异常;数据传输渠道完备,数据不丢失	应用埋点,数据收集能准备充分,保证采集的数据对使用是完备的		埋点符合规范,使得采集的数据规范可用
基础层		任务逻辑正确;集群,计算引擎,调度系统,ETL工具等基础设施保证数据在生产过程能可靠进行.	清洗后的数据在概念上是完整的,对使用是完备的	清洗前后能保证数据在不同时期内概念的连贯性,数据在使用时不会受到业务逻辑变更的影响	格式规范,清晰易懂,对用户友好
应用层	最终结果能在指定时间内产出			多个应用数据能在使用相同的业务概念时,够保证计算口径一致无歧义	同上

数据质量保证策略-时机

影响数据质量的时机:

- 事前: 指数据开发阶段
- 事中: 数据生产进行时,在最终结果预期产出时间之前
- 事后: 数据生产完成,数据到达呈现给用户的阶段

数据质量保证策略-时机

前(规范与验证)	中(发现并恢复)	后(降低影响,运营预期)

数据质量保证策略

前(规范与验证)	中(发现并恢复)	后(降低影响,运营预期)
应用埋点规范	数据生产时效性监控	质量异常数据标记
数据仓库生产规范	表结构变更 发现与确认	数据重导(恢复)工具
业务建模与 指标管理方案	数据异常监控	数据质量问题反馈机制
ETL验证方案和工具	调度系统生产 时运维能力	数据质量评估报表
		数据恢复预期评估

数据安全

- 数据安全等级等级
- 数据脱敏机制
- 统一权限方案与机制

数据安全等级

数据安全等级	分级依据	授权方式
C1(不限制)	适合于公开的数据，并不影响对外发布的数据	直接授权
C2(限制)	不适合对外公开，但是对公司内部人员访问基本无限制	数据接口人负责审批
C3(机密)	适合于部分人可见的数据，丢失或不当使用将显著影响部门开展业务和提供服务等	本部门同C2,跨部门多级审批
C4(绝密)	仅适用于极少部分人可见，信息不安全可能导致公司面临法律或合规的风险	多级审批

数据脱敏机制

- 思路: 通过改变数据形态,降低数据安全等级
- 脱敏原则:
  - 尽可能保留脱敏前数据的业务含义
  - 最大程度上防止黑客进行破解

数据脱敏规则举例

美团网 meituan.com

字段类型	脱敏方法	举例	原则
电话号码	掩码	13812345678-> 13812340000	防止号码泄露，但保留运营商和地区信息
邮件	截断	hxs@163.com -> 6225888e3a1d4a139f5f5db98d846102b2cd0d@163.com	保留邮件域信息
团购密码	加密	402392684339921 9 -> 1298078978	加密后在一定精度上保持唯一性，并与数据类型一致
设备号	加密	ffbacff42826302d9e832b7e907a212a -> b9c2a61972a19bf21b06b0ddb8ba642d	加密后保持唯一性

统一权限方案

- 定义统一的资源表示规范
- 基于RBAC的权限管理模型
- 统一的授权和鉴权接口
- 不同资源类型可定制化资源管理和授权流程界面

统一权限机制

美团网
meituan.com

- 数据资源管理系统
- 数据权限授权与鉴权服务
- 审计系统

大纲

- 数据平台现状
- 数据平台治理服务
  - 资源管理
  - 数据质量
  - 数据安全
- 数据治理系统机制
- 未来的挑战
- QA

治理服务系统机制

美团网
meituan.com

可复用的组件抽象

- 核心组件
- 服务架构

核心组件

- 平台元数据服务 Babel
- 规则引擎 GoldenEye
- 事件订阅与反馈 Delibird
- 通知服务 Spread

服务架构

元数据门户

表

任务

治理服务

数据异常监控

时效性监控

表元数据变更通知与确认

评估

异常数据标记

数据质量报表

引擎

规则引擎 GoldenEye

事件订阅与反馈 Delibird

元数据服务Babel

通知服务 Spread

大纲

- 数据平台现状
- 数据平台治理服务
  - 资源管理
  - 数据质量
  - 数据安全
- 数据治理系统机制
- 未来的挑战
- QA

未来的挑战

- 适应变化,持续完善数据治理系统
- 数据质量评估指标体系
- 提升数据能力的核心:降低数据复杂度

QA

美团网
meituan.com

欢迎提问

xuzhang@meituan.com

美团网

meituan.com

谢谢！